Developing Efficient Code for Decoding AIS Sentences into Useful Datasets

Kiki Beumer

International Association of Marine Aids to Navigation and Lighthouse Authorities

CY Tech Cergy Paris Université

August 8, 2024

Contents

1 Introduction

Automatic Identification System (AIS) data plays a crucial role in maritime navigation and safety. This report assesses the quality and coverage of AIS data within a specified region, focusing on data collected from the Papua New Guinea area between October 28, 2023, and January 26, 2024. The dataset serves as a case study to demonstrate effective decoding of AIS messages, ensuring that the data is well-suited for detailed analysis.

2 Objective

The quantitative analysis of AIS data involves examining the frequency, distribution, and statistical properties of various AIS message types within a specified region. To ensure reproducibility, tools for statistical analysis should be developed.

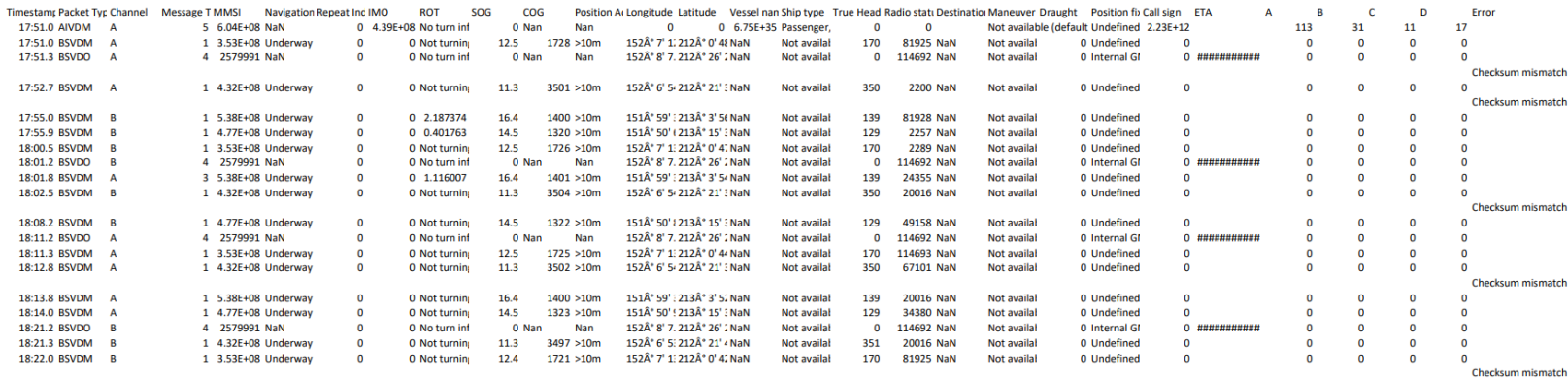
The first chapter focuses on assessing data quality by analyzing distributions, trends, and errors in the AIS messages. This includes checking for inaccuracies such as vessels incorrectly reported as being on land and evaluating positional accuracy. Additionally, a Z-test will be performed to determine whether there is a significant correlation between speed over ground (SOG) and the time intervals between consecutive messages.

In the second chapter, the focus shifts to the geographic coverage of AIS data. The aim is to identify dead zones within the region by comparing actual message frequencies with those predicted based on vessel speed over ground. Extended intervals between messages may indicate areas with poor AIS coverage.

The final chapter employs machine learning techniques to predict quantitative variables from AIS data. This includes predicting speed over ground from location data (longitude and latitude) and forecasting message frequency based on the predicted speed. These predictions will be assessed against actual data to ensure they meet regulatory requirements and accurately reflect vessel behavior. This analysis will identify deviations and serve as a tool for evaluating the quality of the AIS data.

3 Data Types

The subset used for this specific analysis is of size 6.599.735. This data was obtained from 28 October 2023 and 26 January 2024. The csv file looks as follows:



Timestamp	Packet Type	Channel	Message T	MMSI	Navigation	Repeat Ind	IMO	ROT	SOG	COG	Position Ac	Longitude	Latitude	Vessel nan	Ship type	True Head	Radio statu	Destinatio	Maneuver	Draught	Position fi	Call sign	ETA	A	B	C	D	Error
17:51.0	AIVDM	A	5	6.04E+08	NaN	0	4.39E+08	No turn inf		0	Nan	Nan	0	0	6.75E+35	Passenger,	0	0		Not available (default		Undefined	2.23E+12		113	31	11	17	
17:51.0	BSVDM	A	1	3.53E+08	Underway	0	0	Not turnin	12.5	1728	>10m	152Â° 7' 1	212Â° 0' 4	NaN	Not availal	170	81925	NaN	Not availal	0	Undefined	0			0	0	0	0	
17:51.3	BSVDO	A	4	2579991	NaN	0	0	No turn inf		0	Nan	Nan	152Â° 8' 7	212Â° 26' :	NaN	Not availal	0	114692	NaN	Not availal	0	Internal GI	0	##########	0	0	0	0	
																													Checksum mismatch
17:52.7	BSVDM	A	1	4.32E+08	Underway	0	0	Not turnin	11.3	3501	>10m	152Â° 6' 5	212Â° 21' :	NaN	Not availal	350	2200	NaN	Not availal	0	Undefined	0			0	0	0	0	
																													Checksum mismatch
17:55.0	BSVDM	B	1	5.38E+08	Underway	0	0	2.187374	16.4	1400	>10m	151Â° 59' :	213Â° 3' 5	NaN	Not availal	139	81928	NaN	Not availal	0	Undefined	0			0	0	0	0	
17:55.9	BSVDM	B	1	4.77E+08	Underway	0	0	0.401763	14.5	1320	>10m	151Â° 50'	213Â° 15' :	NaN	Not availal	129	2257	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:00.5	BSVDM	B	1	3.53E+08	Underway	0	0	Not turnin	12.5	1726	>10m	152Â° 7' 1	212Â° 0' 4	NaN	Not availal	170	2289	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:01.2	BSVDO	B	4	2579991	NaN	0	0	No turn inf		0	Nan	Nan	152Â° 8' 7	212Â° 26' :	NaN	Not availal	0	114692	NaN	Not availal	0	Internal GI	0	##########	0	0	0	0	
18:01.8	BSVDM	A	3	5.38E+08	Underway	0	0	1.116007	16.4	1401	>10m	151Â° 59' :	213Â° 3' 5	NaN	Not availal	139	24355	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:02.5	BSVDM	B	1	4.32E+08	Underway	0	0	Not turnin	11.3	3504	>10m	152Â° 6' 5	212Â° 21' :	NaN	Not availal	350	20016	NaN	Not availal	0	Undefined	0			0	0	0	0	
																													Checksum mismatch
18:08.2	BSVDM	B	1	4.77E+08	Underway	0	0	Not turnin	14.5	1322	>10m	151Â° 50'	213Â° 15' :	NaN	Not availal	129	49158	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:11.2	BSVDO	A	4	2579991	NaN	0	0	No turn inf		0	Nan	Nan	152Â° 8' 7	212Â° 26' :	NaN	Not availal	0	114692	NaN	Not availal	0	Internal GI	0	##########	0	0	0	0	
18:11.3	BSVDM	A	1	3.53E+08	Underway	0	0	Not turnin	12.5	1725	>10m	152Â° 7' 1	212Â° 0' 4	NaN	Not availal	170	114693	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:12.8	BSVDM	A	1	4.32E+08	Underway	0	0	Not turnin	11.3	3502	>10m	152Â° 6' 5	212Â° 21' :	NaN	Not availal	350	67101	NaN	Not availal	0	Undefined	0			0	0	0	0	
																													Checksum mismatch
18:13.8	BSVDM	A	1	5.38E+08	Underway	0	0	Not turnin	16.4	1400	>10m	151Â° 59' :	213Â° 3' 5	NaN	Not availal	139	20016	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:14.0	BSVDM	A	1	4.77E+08	Underway	0	0	Not turnin	14.5	1323	>10m	151Â° 50'	213Â° 15' :	NaN	Not availal	129	34380	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:21.2	BSVDO	B	4	2579991	NaN	0	0	No turn inf		0	Nan	Nan	152Â° 8' 7	212Â° 26' :	NaN	Not availal	0	114692	NaN	Not availal	0	Internal GI	0	##########	0	0	0	0	
18:21.3	BSVDM	B	1	4.32E+08	Underway	0	0	Not turnin	11.3	3497	>10m	152Â° 6' 5	212Â° 21' :	NaN	Not availal	351	20016	NaN	Not availal	0	Undefined	0			0	0	0	0	
18:22.0	BSVDM	B	1	3.53E+08	Underway	0	0	Not turnin	12.4	1721	>10m	152Â° 7' 1	212Â° 0' 4	NaN	Not availal	170	81925	NaN	Not availal	0	Undefined	0			0	0	0	0	
																													Checksum mismatch

Figure 1: Output csv file

The following data types have been extracted from the AIS messages:

Table 1: AIS Data Types

Columns in *.csv file	**Description**	**Data Type**	**Example**
Timestamp	Timestamp from the AIS basestation	datetime	31/12/2015 23:59:59
Packet Type	Defines the packet type, which can be defined in different formats	string	AIVDM
Channel	Radio channel code	string	A
Message Type	Specifies the kind of information being transmitted	float	1
MMSI	MMSI number of the vessel	string	538009733
Navigation Status	Navigational status from AIS message if available	string	Under way using engine
Repeat Indicator	Message repeat count	float	0
IMO	IMO number of the vessel	float	9675626
ROT	Rate of turn in degrees/min	string	2.1873
SOG	Speed over ground in knots	string	12.5
COG	Course over ground relative to true north	float	1322
Position Accuracy	Accuracy of the DGPS-quality fix	string	¿10m
Longitude	Geographic coordinate	float	37.7749° E
Latitude	Geographic coordinate	float	37.7749° N
Region	Region based on IARU Locator grid system	string	QH
Vessel name	Name of vessel	string	3.17395E+35
Ship Type	Type of vessel	string	Cargo, all ships of this type
True Heading	Heading of the vessel in degrees	string	170
Radio status	State of the vessel's AIS transceiver	float	81925
Destination	Intended port of arrival or destination of the vessel	float	...
Maneuver Indicator	Vessel's current maneuvering status	string	Special maneuver
Draught	Vessel's vertical distance between the waterline and the bottom of the hull in meters	float	17.5
Position Fixed type	Type of positional fixing device from the AIS message	string	Internal GNSS
Call sign	unique identifier assigned to a vessel's radio communication equipment	float	...
ETA	Estimated time of arrival	datetime	10/28/2023 7:12:00 AM
A	Length from GPS to the bow in meters	float	202
B	Length from GPS to the stern in meters	float	33
C	Length from GPS to starboard side in meters	float	13
D	Length from GPS to port side in meters	float	25
Error	Any error messages that appear	string	Checksum mismatch

The distribution of the following quantitative data types was studied:
- ROT (rate of turn)

- SOG (speed over ground)
- COG (course over ground)
- Draught
- A (length from GPS to the bow in meters)
- B (length from GPS to the stern in meters)
- C (length from GPS to starboard side in meters)
- D (length from GPS to port side in meters)

Based on the information available in the "Message Type" column, we will drop rows with 'NaN' values to obtain a more informative subset. The presence of different message types results in a significant number of 'NaN' values, which will be removed for clearer analysis.

4 Quantitative Analysis per Region

The following code can be used to reproduce the histogram and observe the distribution of the quantitative variables. The lineplot through the histogram is the KDE. This is a non-parametric way to estimate the probability density function (pdf) of a random variable.

```
# Exchange 'Draught' with any other column name if needed
ais_draught = ais.dropna(subset='Draught')
ais_draught = ais_draught[ais_draught['Draught'] != 0.0]

sns.set_theme(style="darkgrid")
sns.histplot(data=ais_draught, x="Draught", kde=True)
plt.show()
```

Listing 1: Code Histogram and KDE

4.1 Speed over Ground

In case `SOG == 102.3`, this is the numerical equivalent of 'NaN' for this variable. These values are dropped. The resulting size of the sample is 4.597.463. KDE shows that this distribution is approximately normal.

To closely examine the speed over ground in a specific area, we can subset the data for a particular Maidenhead region; QH69. This region is more precisely defined by the two numbers following the two capital letters that identify the general area.

```
#Subset for a region
ais_sog = ais_sog[ais_sog["Region"]=='QH69']

#Create the histogram
sns.set_theme(style="darkgrid")
sns.histplot(data=ais_sog, x="SOG", kde=True)
plt.xlim(0, 25)
plt.show()

#Mean, SD, min ,max
```

```
print(ais_sog["SOG"].mean())
print(ais_sog["SOG"].std())
print(ais_sog["SOG"].min())
print(ais_sog["SOG"].max())

# OUTPUT:
# Mean: 12.386505500958968
# SD: 2.4601595797148024
# Min: 0.0
# Max: 99.6
```

Listing 2: Subsetting for region

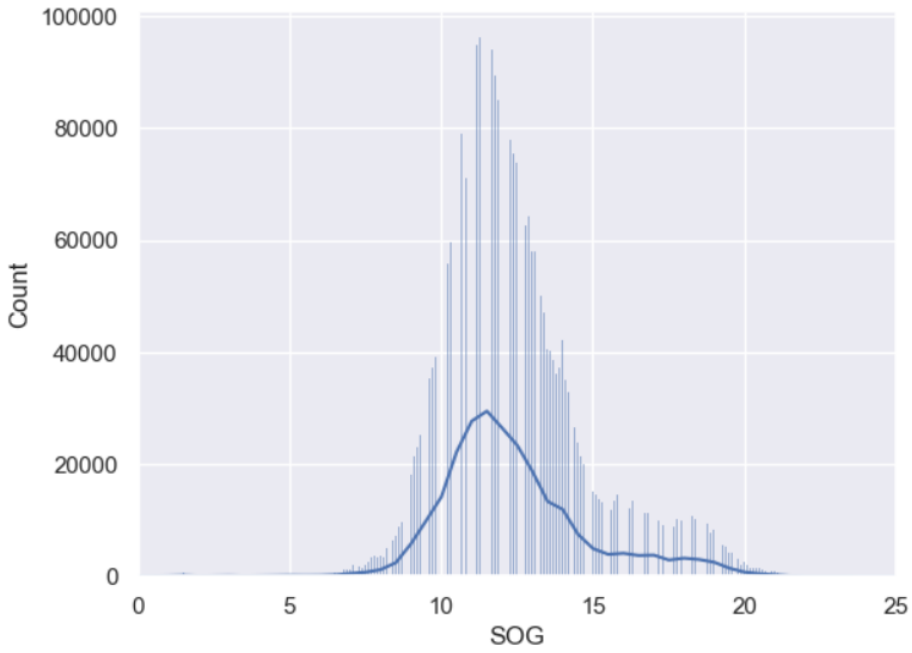


Figure 2:
Speed over ground distribution for region QH69

Once we have subsetted the data, we can plot the SOG again. This detailed view can help identify regions where there might be other abnormal activity. We can also combine this data with the transmitting frequency to determine whether the requirements are being met or to identify potential dead zones. Region QH69 shows no significant deviations from the original dataset. The mean and distribution appear consistent.

4.2 Course over Ground

In case `COG == 3600`, this is the numerical equivalent of 'NaN' for this variable. These values are dropped. The resulting size of the sample is 4.594.227. The histogram shows a trimodal distribution.

We can repeat the same subsetting method for region QH69 as we did with SOG. The resulting histogram is aw follows:

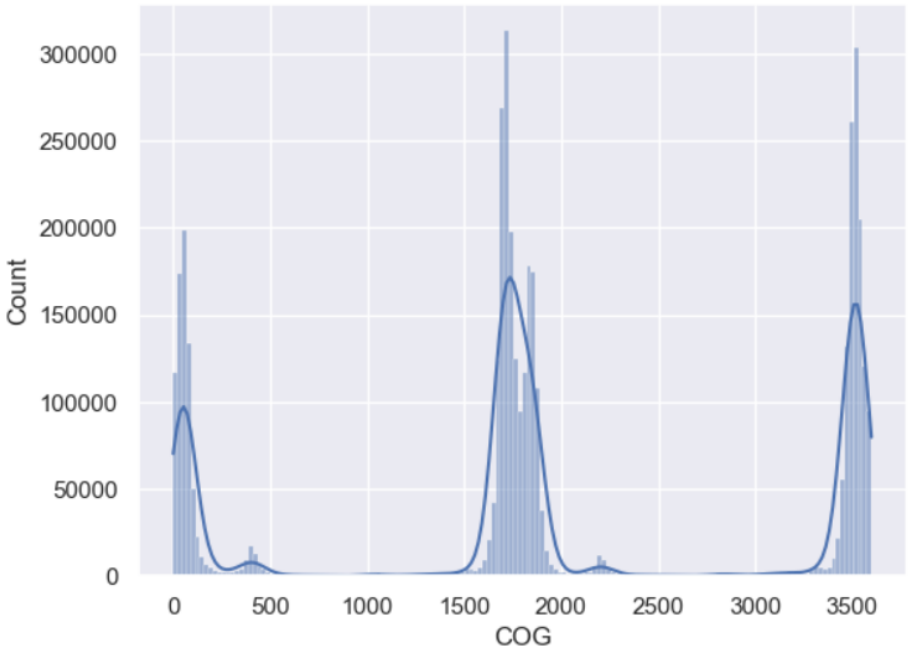


Figure 3:
Course over ground distribution for region QH69

Again, we see the same distribution as the complete dataset.

```
# OUTPUT:
# Mean: 1982.484944691728
# SD: 1215.848305429338
# Min: 0.0
# Max: 3599.0
```

Listing 3: Output COG in region QH69

4.3 Draught

The resulting size of the sample after dropping 'NaN' values is 117.568.

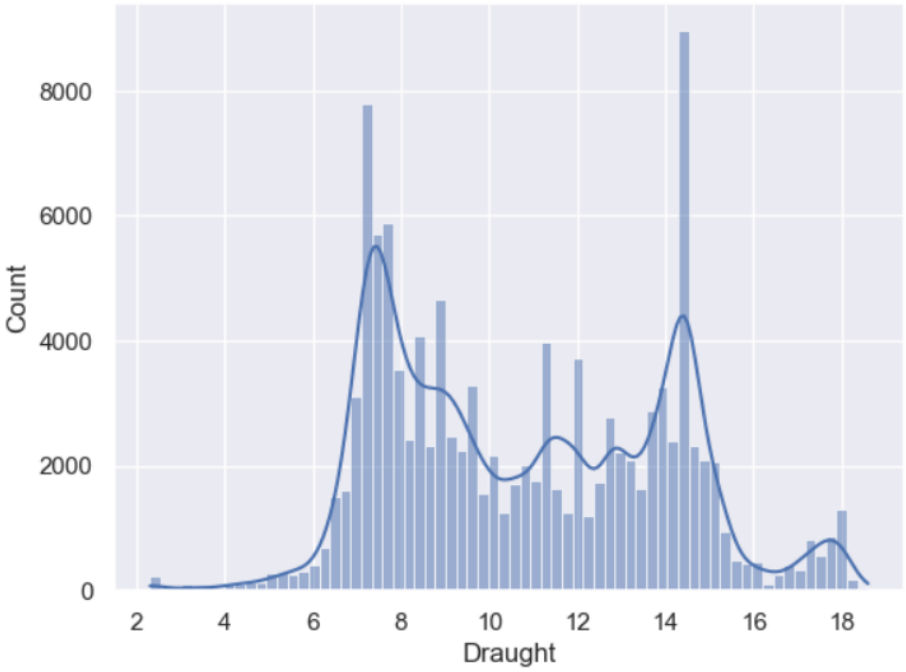


Figure 4:
Draught distribution

```
print(ais["Draught"].mean())
print(ais["Draught"].std())
print(ais["Draught"].min())
print(ais["Draught"].max())

# OUTPUT:
# Mean: 0.19248814289353253
# SD: 1.4908421152141382
# Min: 0.0
# Max: 18.6
```

Listing 4: Output COG in region QH69

4.4 Vessel's length

All rows where `A==0` are dropped, this is the numerical equivalent of 'NaN' for this variable. The resulting size of the sample is 118.579. All length are displayed in one plot.

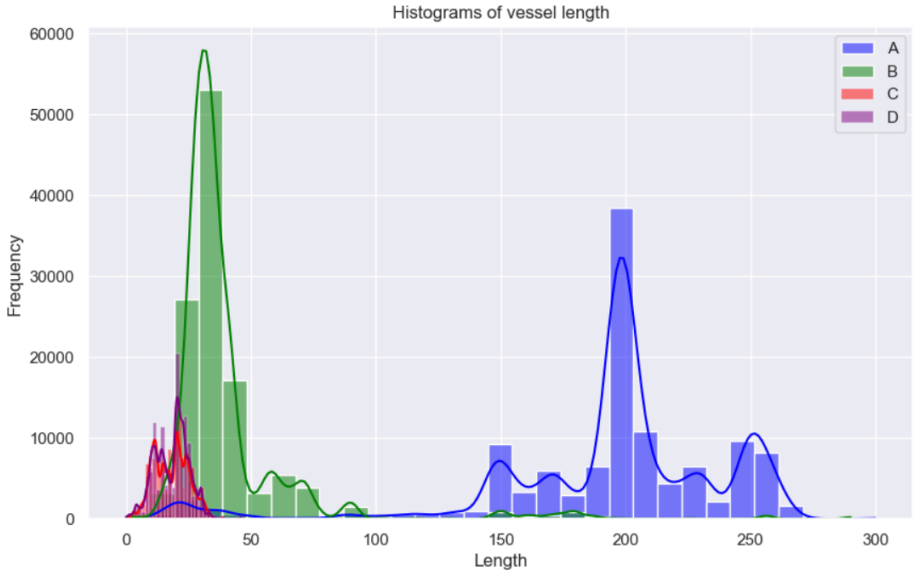


Figure 5:
Vessel's length distribution

4.5 Unique vessels

Refers to the count of distinct vessels that are identifiable within the dataset based on their unique identifiers (Maritime Mobile Service Identity (MMSI) number). Each vessel is represented as a unique entity, and counting the number

of unique ships provides insight into the diversity and volume of maritime traffic captured by the dataset. With a simple `.nunique()` function we can get that this dataset is provided by 1973 unique vessels.

```
# Extract date from the timestamp
ais_cleaned['date'] = ais_cleaned['Timestamp'].dt.date

#Create table
unique_vessels = df.groupby('date')['MMSI'].nunique().reset_index()
unique_vessels.columns = ['Date', 'Unique vessels']
```

Listing 5: Get Unique vessel code

To visualize the changes in traffic volume over time, we first create a pivot table or use a groupby operation. Afterward, we plot this aggregated data. This allows us to observe that there is fluctuation in the number of vessels sending AIS messages each day. However, there is no specific trend to be seen. **With help of ML we could predict the number of vessels traveling each day.**

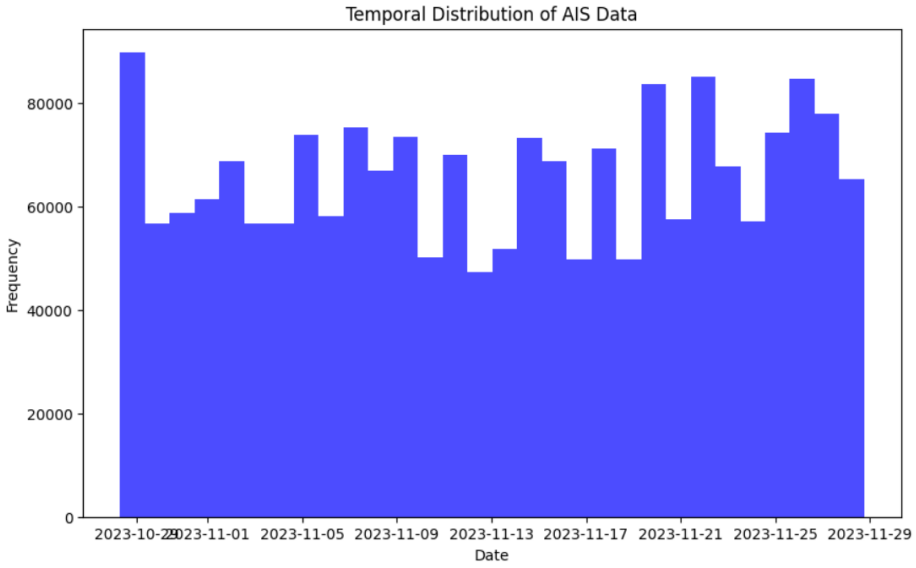


Figure 6: This map shows the IARU Grid Locator system, overlaid on a World Map

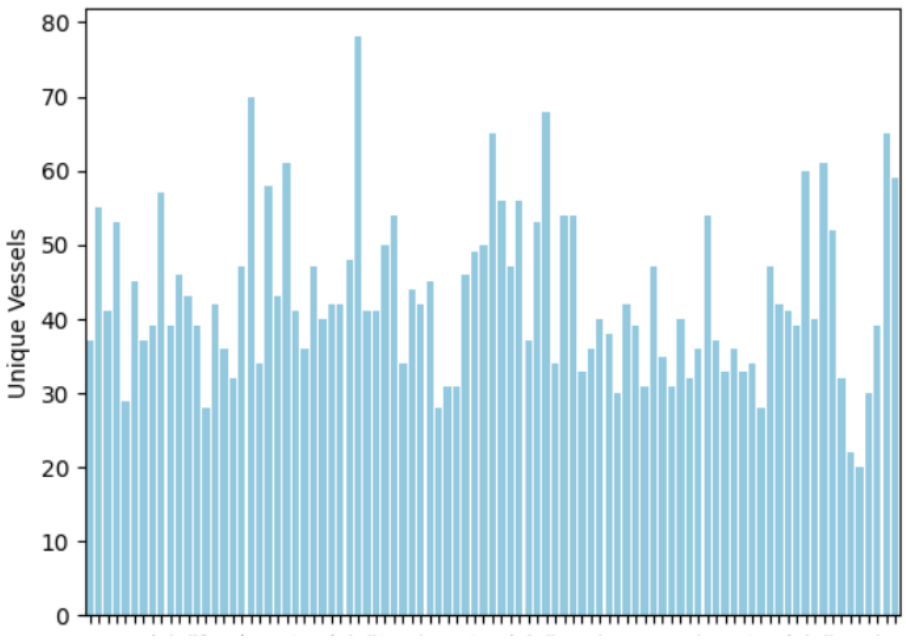


Figure 7: Unique vessle over time

~~4.6 Data Types~~

Table 2: AIS Data Types

	Length from GPS to the bow	Length from GPS to the stern	Length from GPS to starboard side	Length from GPS to port side
	Size A	**Size B**	**Size C**	**Size D**
Mean	193.77	40.15	18.0	18.42
SD	48.32	27.59	6.88	6.70
Min	9.0	1.0	1.0	1.0
Max	300.0	290.0	39.0	46.0

5 Data Quality in Region

Refers to analyzing the frequency or detail level of position reports for individual trips made by ships. This involves assessing how often position updates are recorded for each voyage undertaken by a ship. This includes the previously performed analysis as well, such as the geographical boundary and the temporal distribution analysis. To begin, we assess the dataset for errors by identifying entries with a "Checksum mismatch." Additionally, we evaluate the percentage of data points that have a position accuracy exceeding 10 meters.

5.1 Error checking

The provided code was designed to count the number of error messages within the dataset. However, no error messages were detected in this dataset. This outcome is attributed to the fact that proprietary AIS messages do not generate explicit error messages but instead utilize the "Datafield" columns. Additionally, our analysis did not reveal any mismatched checksums or unexpected values in the data during decoding.

```
ais[ais["Error"].notna()]["Error"].value_counts()

#OUTPUT:
# 0
```

Listing 6: Python code for main

5.2 Position Accuracy

When we count the values in the "Accuracy Position" column, we find that 65.14% of the total data has a position accuracy greater than 10 meters. It's important to note that position accuracy is not always provided, depending on the message type. Among the available position accuracy values, 93.48% of the data has a position accuracy greater than 10 meters.

Table 3: AIS Data Types

Position Accuracy	**0**	**<10m**	**>10m**
MMSI			
0	163849	0	0
125962272	0	0	38
209034000	0	0	1963
209669000	0	0	1994
209691000	0	0	43
...	...	...	...
636093180	0	1916	0
636093157	0	0	34
6360931800	0	130	1732

7,54% vessels of the total messages were able to transmit AIS messages with a position accuracy smaller than 10 meters. Which off course means that 92,46% of the vessels only transmitted messages with their position accuracy exceeding 10 meters.

5.3 On land values

```
# pip install global_land_mask
from global_land_mask import globe
```

```
3
4 # Apply the globe.is_land function row-wise to create the 'Is Land'
      column
5 ais['Is Land'] = ais.apply(lambda row: globe.is_land(row['Latitude'],
      row['Longitude']), axis=1)
6
7 ais.head()
```

Listing 7: Python code for main

Creating a count table of these values reveals that, in this subset of the dataframe, there are no instances where a vessel was recorded as being on land. While this observation might be useful for identifying potential errors, it cannot be conclusively interpreted, as there could be other explanations for this result.

```
1 island_table = ais['Is Land'].value_counts().reset_index()
2 island_table.columns = ['Is Land', 'Count']
3
4 island_table
```

Listing 8: Python code for main

Table 4: AIS Data Types

Is Land	**Count**
False	118.899
True	0

This code produces a new column in the dataset *ais*; "Is Land". This column gives the boolean value "True" is the longitude and latitude point were on land. Similarly the value "False", indicates that this vessel was on water when transmitting the message.

6 Predict message frequency from SOG

First, we will look at the distribution of the time frequency between consecutive AIS messages. Created this extra column is easily done with a *.diff* function between consecutive rows, since the data is already sorted by timestamp. The *pd.infer_freq* function shows that the time between consecutive messages of a vessel is not constant. Then plotting this data shows that the time interval has a right skewed distribution.

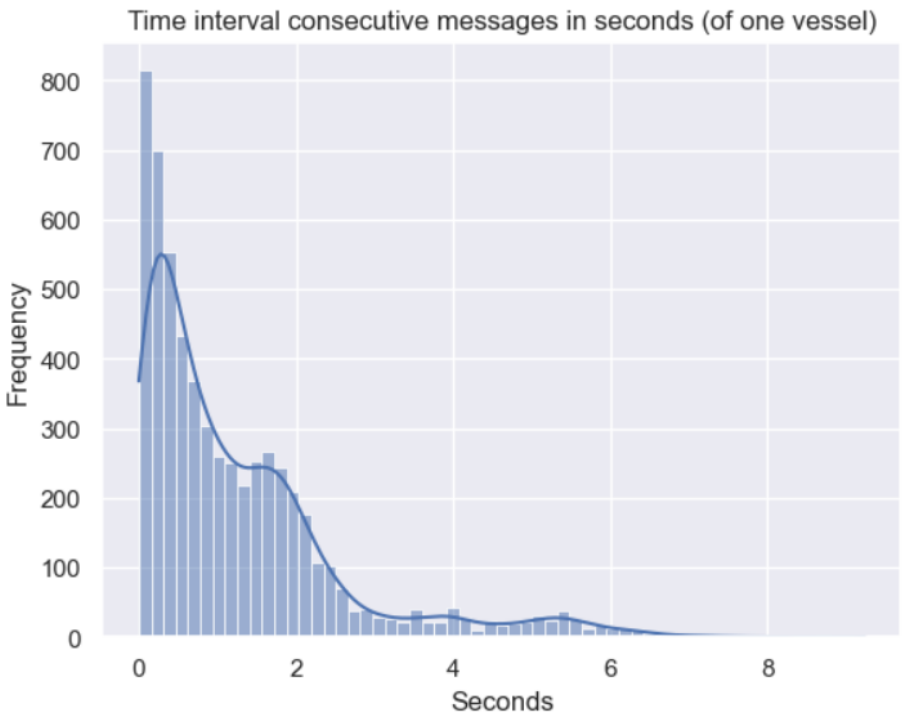


Figure 8: Time interval frequncy

According to established guidelines, there should be a correlation between the time intervals of AIS transmissions and the Speed Over Ground (SOG). We should be able to see a positive correlation between the speed over ground and the time intervals between consecutive messages.

6.1 Correlation between SOG and Message Frequency

We calculated the correlation coefficient between *ais["time_diff"]* and *ais["SOG"]*.

```
correlation = ais["time_diff"].corr(ais["SOG"])
correlation

#OUTPUT: -0.0001697560829245 13
```

Listing 9: Python code for correlation calculation

The resulting correlation coefficient is 0.00017, indicating an almost negligible correlation between the time intervals of AIS transmissions and the SOG. This lack of correlation suggests one of two possibilities: either the rules regarding transmission intervals and SOG are not being adhered to, or there are gaps or "dead zones" in the data that are affecting the results. Further investigation is required to determine the cause of this discrepancy. The scatterplot below however, does not show this.

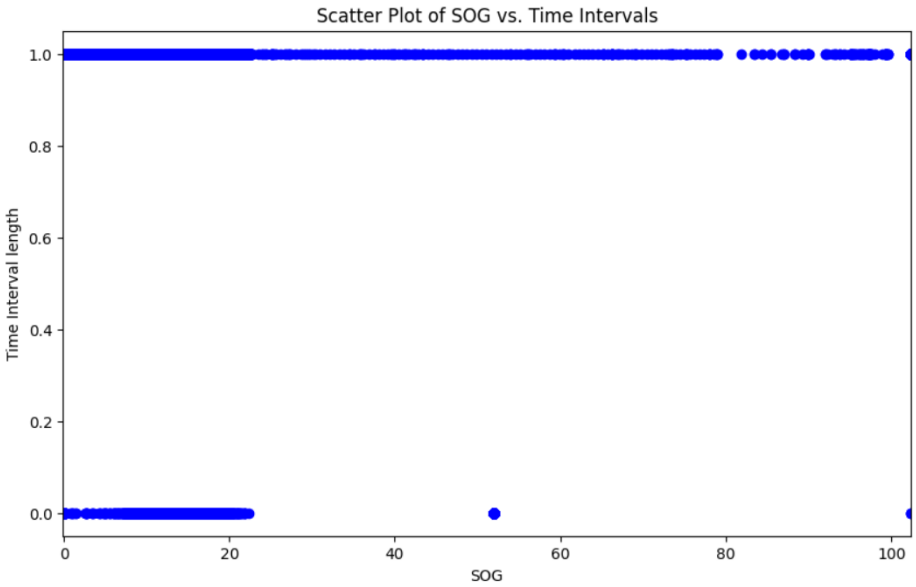


Figure 9:
Scatterplot Time interval length versus SOG

Predict the Message frequency from the SOG with ML;

7 Geographical boundaries and coverage

7.1 Objective

This refers to the defined spatial limits within which the data was collected or is relevant. For AIS data, this involves specifying the latitude and longitude ranges that encompass the area covered by the dataset. This includes identifying the maximum and minimum longitude and latitude values, as well as checking for any missing values and identifying dead zones.

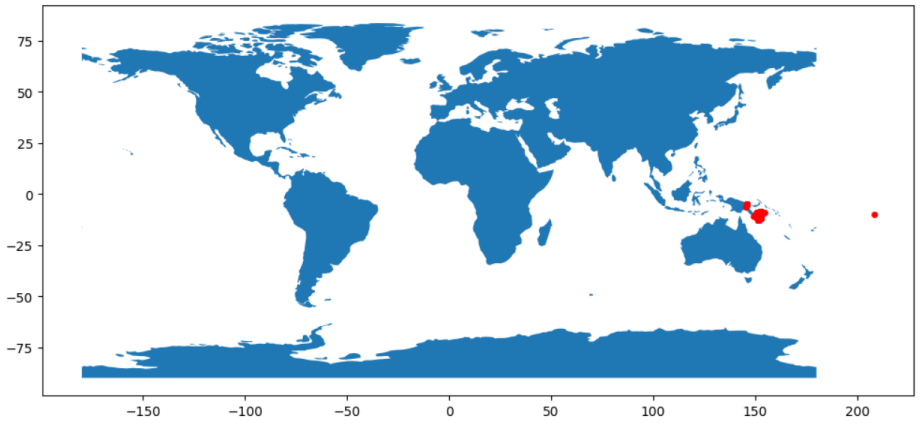


Figure 10:
Geopandas map of datapoints

This map would be more suitable if the dataset would contain points of more than one IARU Maidenhead region. However, it is helpful to visualize the location of the data points.

When zooming in to the region, we can spot the dead zone areas. *how to recover these points???*

7.2 Limitations

To observe the geographical boundaries we can display the longitude and latitude in a scatter plot. The dotted red lines are the minimum and maximum longitude and latitude in the dataset. We can use a scatterplot to visualize the coverage of longitude and latitude within these boundaries.

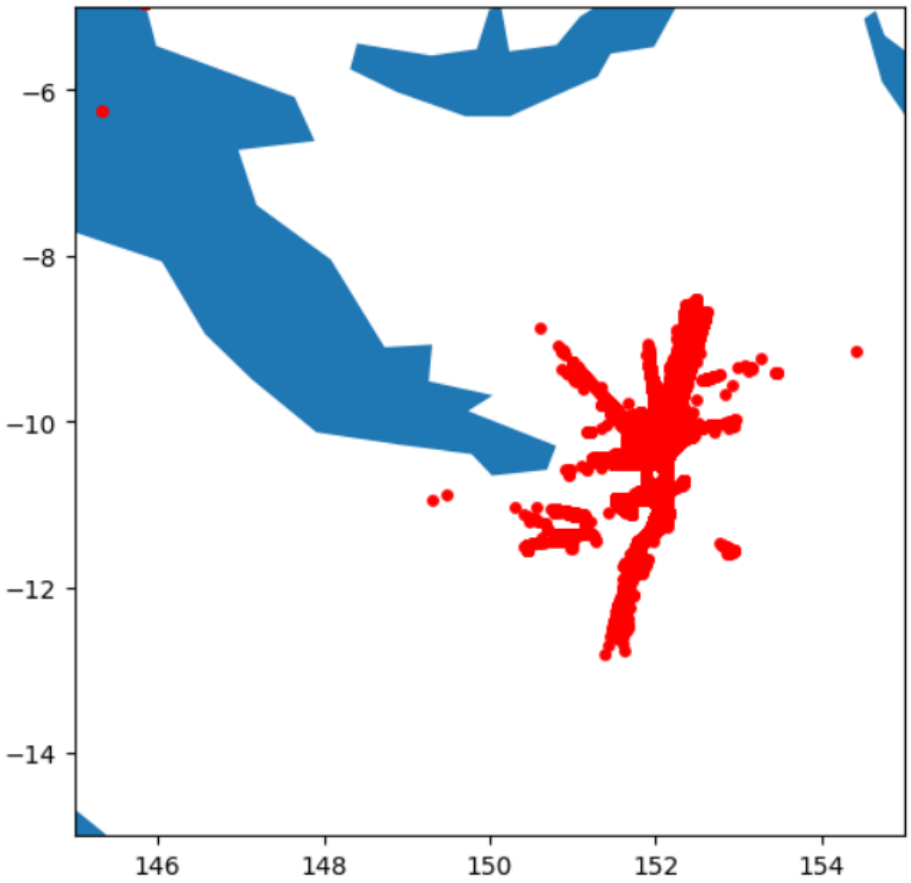


Figure 11:
Geopandas map of datapoints

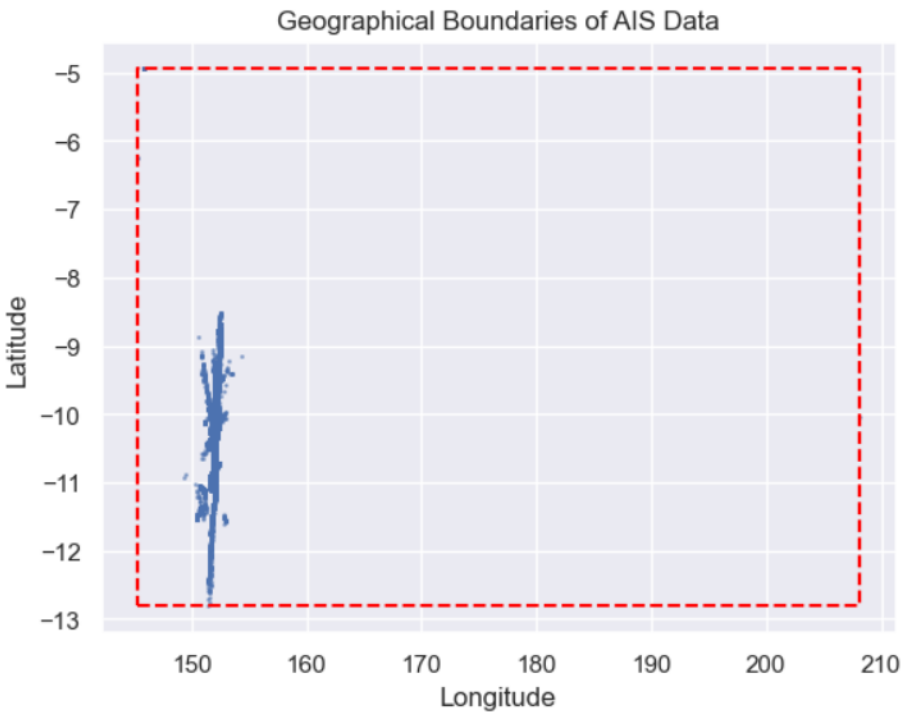


Figure 12: Geographical coverage of dataset

However, this tool is not effective for identifying missing values, as the plot also includes longitude and latitude points that fall on land.

8 Predict Quantitative variables

9 Conclusion

Add when finished

10 References

Danish Maritime Authority - AIS. Retrieved from:
https://github.com/dma-ais
Automatic identification system. Retrieved from:
https://en.wikipedia.org/wiki/Automatic_identification_system
AIS data Danish Maritime Authority. Retrieved from:
https://www.dma.dk/safety-at-sea/navigational-information/ais-data
GH AIS Message Format. Retrieved from:
https://www.iala-aism.org/wiki/iwrap/index.php/GH_AIS_Message_Format
AIVDM/AIVDO protocol decoding. Retrieved from:
https://gpsd.gitlab.io/gpsd/AIVDM.html
pyais. Retrieved from:
https://github.com/M0r13n/pyais/tree/master?tab=readme-ov-file
ais-protocol-decoding. Retrieved from:
https://github.com/doron2402/ais-protocol-decoding
AIRU Maidenhead Grid Locator Retrieved from:
https://www.mapability.com/ei8ic/maps/gridworld.php